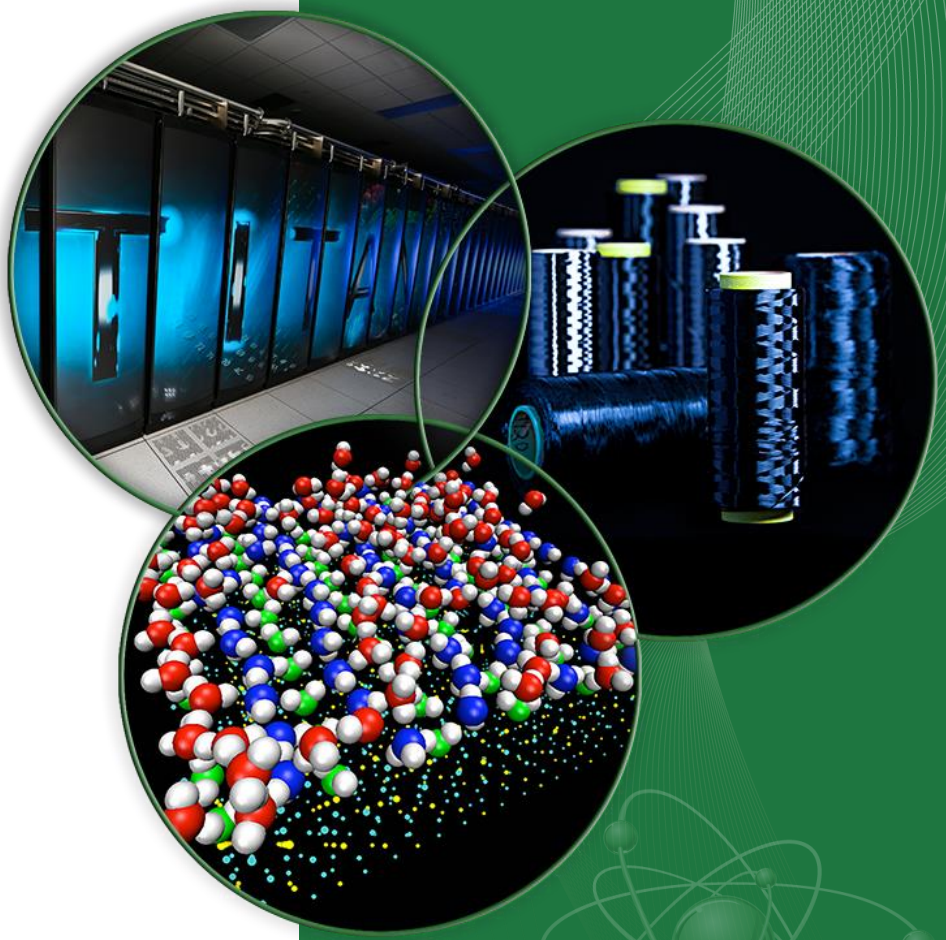
Highlights from Beyond CMOS Computing Workshops



Presented at the
Extreme Heterogeneity Workshop

Neena Imam
Director of Research Collaboration
Computing and Computational Sciences

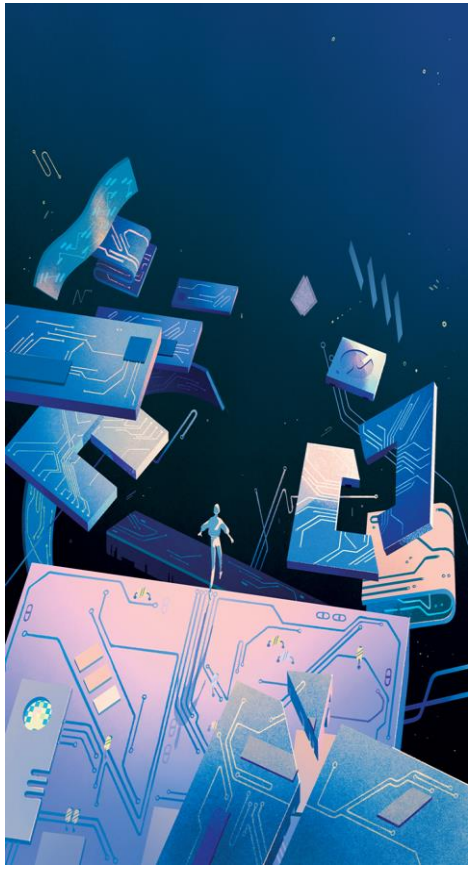
January 24, 2018

ORNL is managed by UT-Battelle
for the US Department of Energy

OAK RIDGE National Laboratory

President Obama's FY17 budget request mentions *Beyond Moore's Law*

- As noted in the NSCI, the era of silicon-based microchips advancing in accordance with Moore's Law ….is nearing an end due to limits imposed by fundamental physics. ASCR will invest $12 million across research and facilities to understand the impacts these technologies may have on our applications. Beginning in FY 2017, the computer science and computational partnerships activities will invest $7 million to initiate new research efforts on technologies "Beyond Moore's Law," responding to the NSCI and recommendations made by the Secretary of Energy Advisory Board, to understand the challenges that these dramatically different technologies pose to DOE mission applications and to identify the hardware, software and algorithms that will need to be developed for DOE mission applications to harness these developing technologies.



OAK RIDGE National Laboratory

Context of the Beyond CMOS Computing workshops

- ORNL-DoD jointly hosted two *Beyond CMOS Computing* workshops in 2016 and 2017

  Beyond CMOS Computing: The Interconnect Challenge (2017)
  http://beyondcmos.ornl.gov/

  Looking Beyond CMOS Technology for Future HPC (2016)
  http://beyondcmos.ornl.gov/2016/

- Workshop organizers

Co-Chairs	DoD Steering Committee
Neena Imam, ORNL	Janice Elliott
Barney Maccabe, ORNL	John Bolger
Jeff Nichols, ORNL	Kristin Lucas

- Motivation

  Review of beyond CMOS landscape

  Educate the attendees (~ 50% of attendees from the intelligence community)

  Workshop not directly tied to any funding opportunities

Organizations represented at the *Beyond CMOS Computing w*orkshops

- Oak Ridge National Laboratory (ORNL)
- Pacific Northwest National Laboratory (PNNL)
- MIT Lincoln Laboratory
- Department of Defense (DoD)
- Department of Energy (DOE)
- National Science Foundation (NSF)
- National Institute of Standards and Technology (NIST)
- Institute of Defense Analysis (IDA)
- Intelligence Advanced Research Projects Activity (IARPA)
- Office of Naval Research
- Scientific Research Corporation
- SGI, Mellanox , IBM
- Northrop Grumman Corporation
- D-Wave
- Booz Allen Hamilton
- Natural Intelligence
- University of Missouri Kansas City, UCLA, Stanford, Georgia Tech, NC State, Florida State, University of Lyon, Notre Dame
- Mayo Clinic

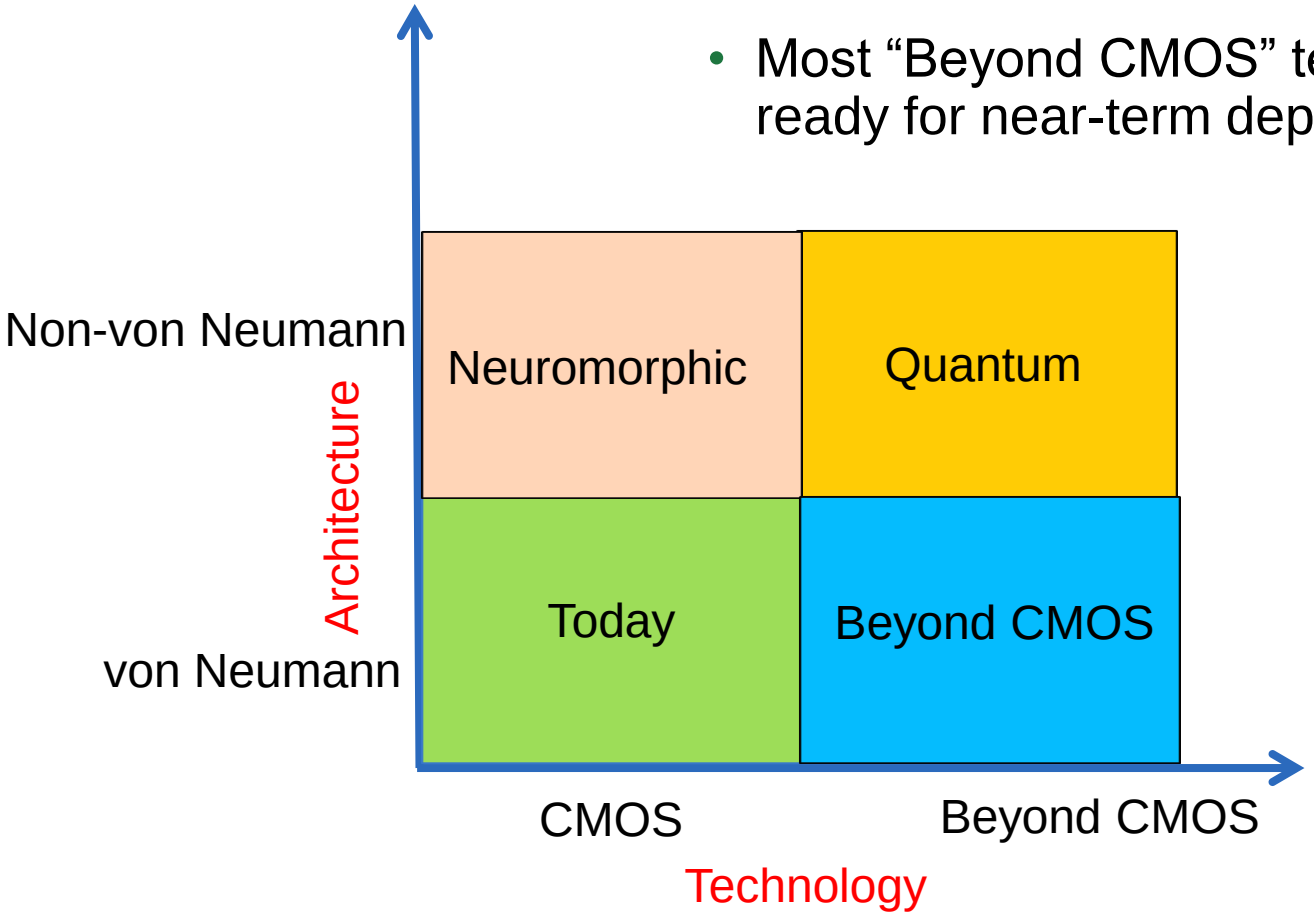
OAK RIDGE National Laboratory

Looking Beyond CMOS Technology for Future HPC, 2016

- ORNL-DoD jointly hosted this workshop on April 5-6, 2016 in Maryland.
- Four sessions based on the following focus areas: nanomaterials, quantum computing, superconducting computing, and emerging processor and memory architectures.
- Other areas were not included due to time/resource limitations: neuromorphic computing, probabilistic computing, optical computing……
- This workshop was a status report on the Beyond CMOS landscape.

End of Moore's Law will spur innovation and lead to new architectures

- We still use von-Neumann architecture
- Sustained growth of transistor technology allowed us to ignore architecture so far
- Most "Beyond CMOS" technology are not ready for near-term deployment..

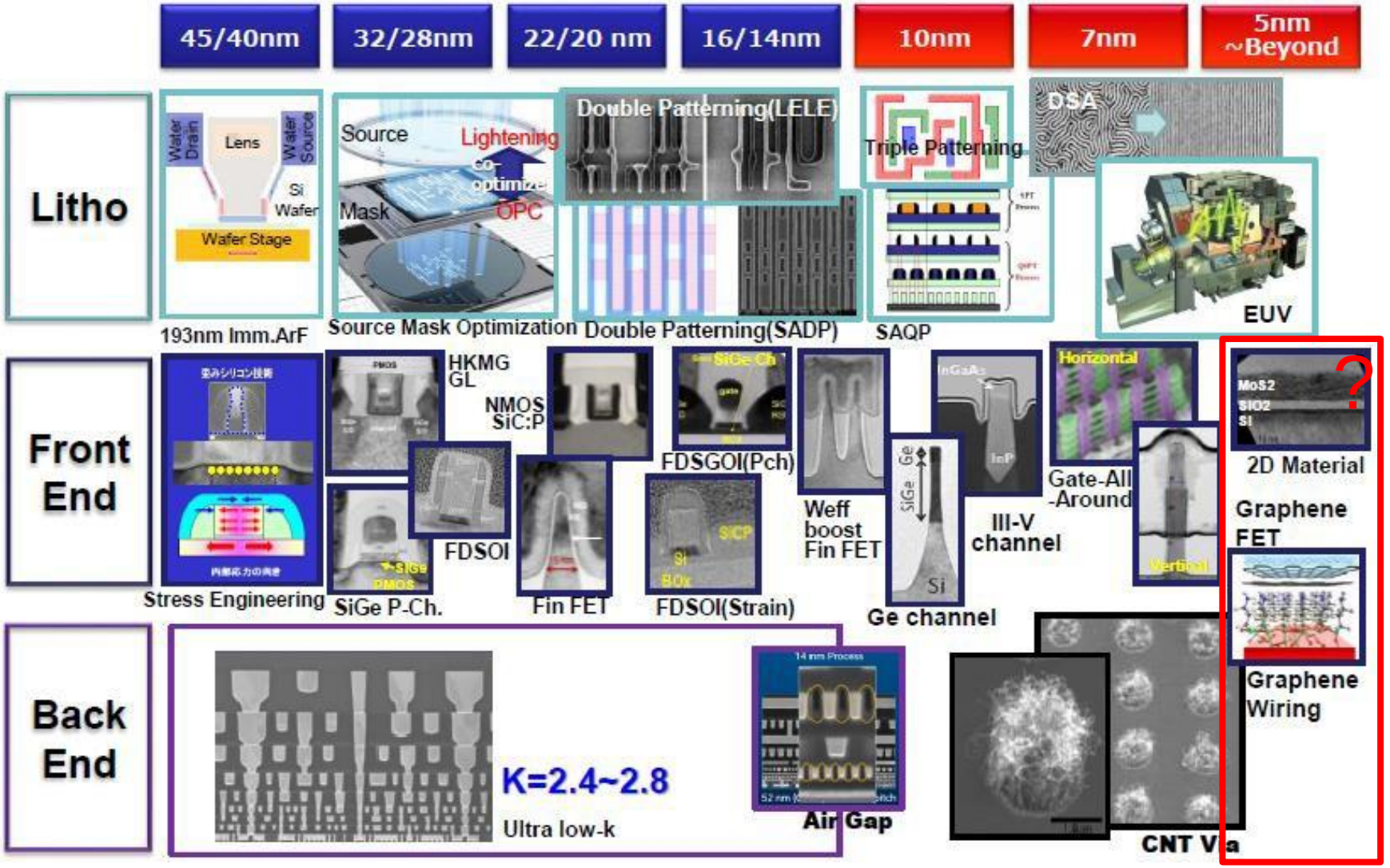


Focus areas of the 2016 workshop

Session 1: Nanomaterials for Future HPC

- New materials might hold the answer for beyond CMOS technology for better energy efficiency and smaller footprint
- Most promising: carbon based nano-materials such as graphene, carbon nanotube based transistors
- Challenges:
  - Move from materials to systems
  - Fabrication processes need to be improved
  - Maintaining advantageous properties of nanomaterials at very large scale

The International Technology Roadmap for Semiconductors (ITRS)



Courtesy of: Yuzo Fukuzaki, cited from M. Badaroglu, "More Moore scaling: opportunities and inflection points," ERD Meeting: Bridging Research Gap between Emerging Architectures and Devices, Feb 27, 2015

Advanced FETS

Silicon and Silicon Germanium FinFETs will extend CMOS to 10nm node and likely 7nm node

Taller and tighter fin pitch FinFETs are low risk path to 5nm node, but short channel effects are a concern

Gate all Around (GAA) FETs are 2nd option for 5nm node, but stacking wires is a challenge

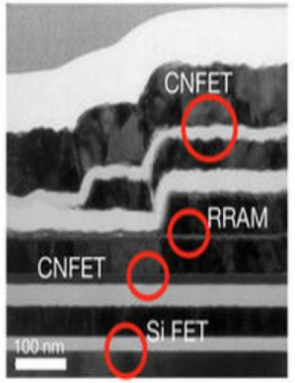
Band structure engineered Germanium and III-V QW FinFETs show great performance advantage, but manufacturability is a challenge

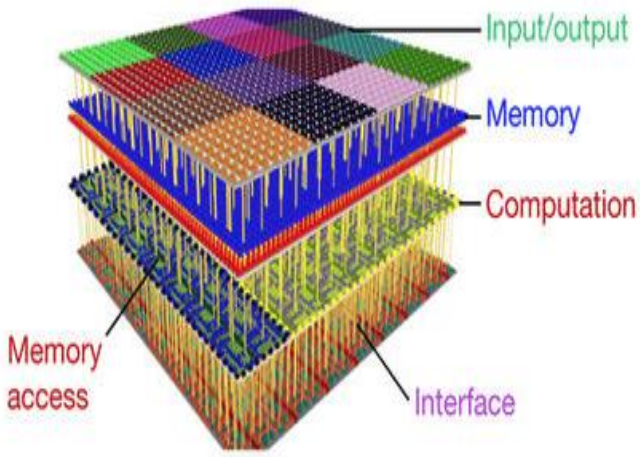
Band engineered Tunnel FETs may augment CMOS to support heterogeneous core processors, but experimental demonstration of steep slope with acceptable drive current is still elusive

Credit: Suman Datta (Notre Dame), Beyond CMOS Technologies for Future HPC, 2016

Carbon NanoTube (CNT) processor N3XT

- N3XT: enabled by emerging nanotechnologies
  - Energy efficient logic (e.g., CNFETs)
  - On-chip nonvolatile memories (e.g., RRAM)
  - Densely-connected logic ↔ memory (e.g., ILVs)
- How can these benefit HPC workloads?





Credit: MIT

Instead of relying on silicon-based devices, researchers and Stanford University and MIT have built a new chip that uses carbon nanotubes and resistive random-access memory (RRAM) cells. The two are built vertically over one another, making a new, dense 3-D computer architecture with interleaving layers of logic and memory. This work was funded by DARPA, NSF, SRC, STARnet SONIC, and member companies of the Stanford SystemX Alliance.

Focus areas of the 2016 workshop

Session 2: Quantum Computing

- Quantum computers could offer a giant leap in speed—but only for certain applications.
- Challenges:

  Qubit materials design: The quantum computing community has yet to settle on a single qubit technology.

  Characterization and modeling of qubits: Most powerful supercomputers of today cannot model a quantum computing system containing more than ~50 qubits.

  Development of balance of system components: development of quantum-grade electrical, optical, and mechanical connections, cryogenic instrumentation, thermal and heat transfer management, and modeling of cryogenic fluid flows and heat transfer.

  Quantum – Classical programming models: programming models and programming languages for computations with both quantum and classical components.

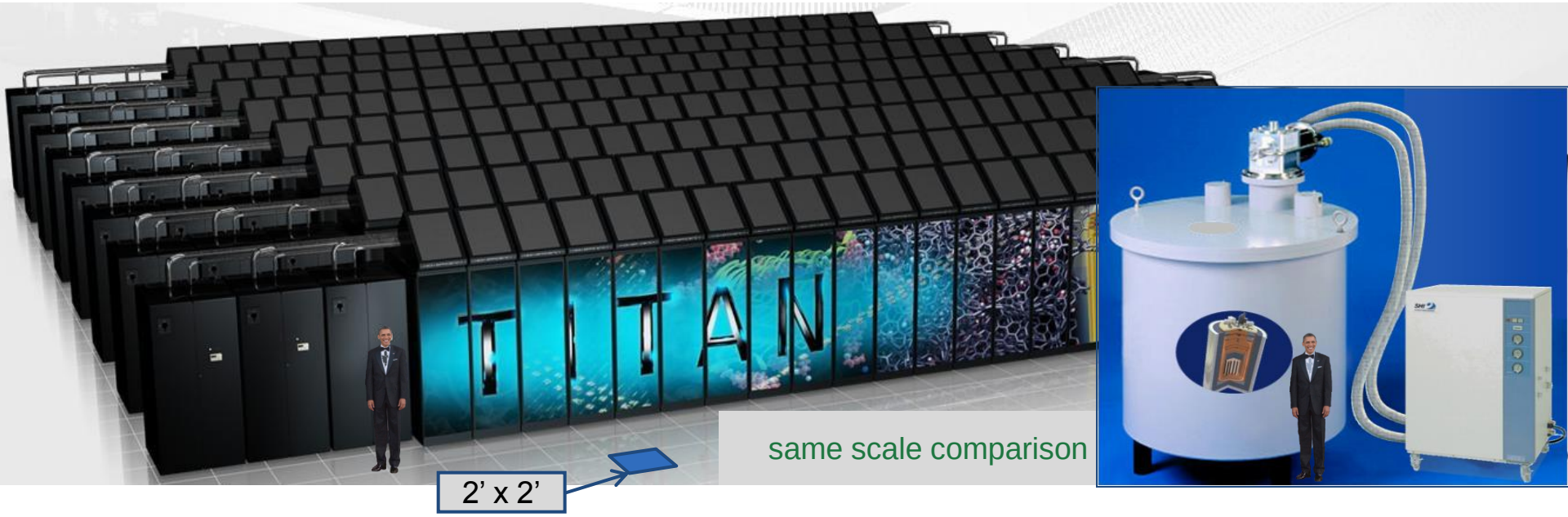
Focus areas of the 2016 workshop

Session 3: Superconducting Computing

- Advantages are blazing speed and very low power
- Challenges:
  - The superconducting processors work at cryogenic temperature and this additional requirement of refrigeration imposes systems engineering challenges
  - Cryogenic memory capacity is presently limited to low density
  - Architectural solutions for signal amplification are needed to realize a superconducting computer
  - The CAD design tools need to be improved to a point where logic designers are able to design logic chips without Ph.D. level knowledge of the physics of superconductivity
  - Cryogenic interconnects…more on this later

Focus areas of the 2016 workshop

System Comparison (~20 PFLOP/s)



	Titan at ORNL	Superconducting Supercomputer	
Performance	**17.6** PFLOP/s	**20** PFLOP/s	~1x
Memory	**710** TB (0.04 B/FLOPS)	**5** PB (0.25 B/FLOPS)	7x
Power	**8,200** kW avg. (not included: cooling, storage memory)	**80** kW total power (includes cooling)	0.01x
Space	**4,350** ft^2 (404 m^2, not including cooling)	~**200** ft^2 (includes cooling)	0.05x
Cooling	additional power, space and infrastructure required	All cooling shown	

Credit: Marc Manheimer (IARPA), Beyond CMOS Computing: The Interconnect Challenge, November, 2017

OAK RIDGE National Laboratory

Focus areas of the 2016 workshop

Key Factors

- Approach based on:
  - ***Near-zero energy*** superconducting interconnect
  - ***New*** SFQ logic with no static power dissipation
  - ***New energy efficient*** cryogenic memory ideas

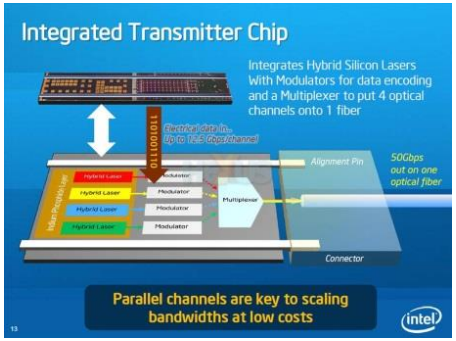
  IARPA C3 program basis
  - ***Optical*** ingress/egress
  - ***Commercial*** cryogenic refrigerators

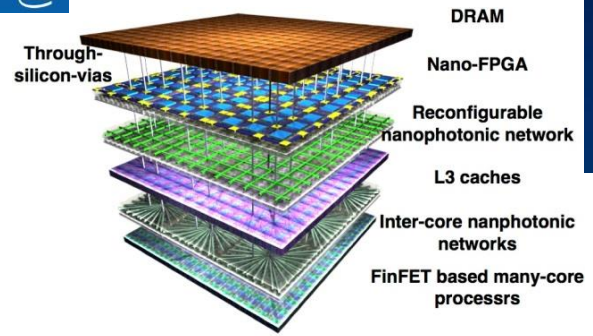
Credit: Marc Manheimer (IARPA), Beyond CMOS Computing: The Interconnect Challenge, November, 2017

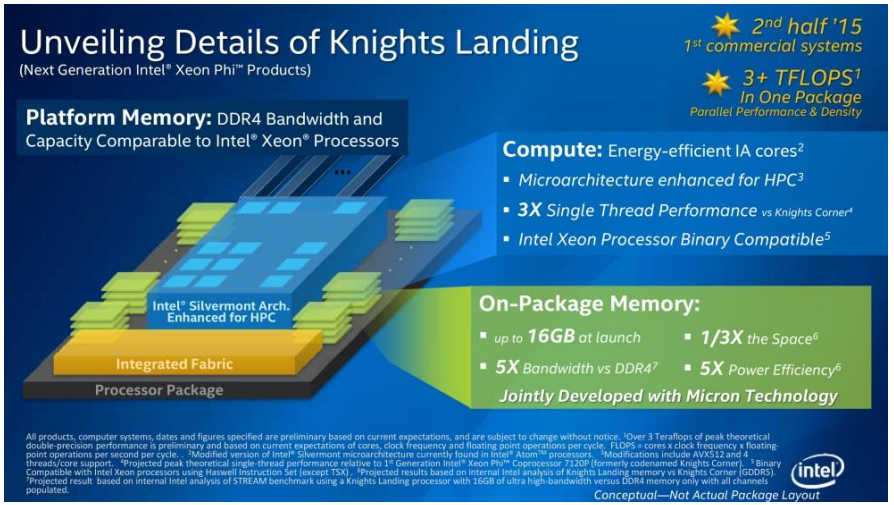
Focus areas of the 2016 workshop

Session 4: Emerging Processor and Memory Architectures

- Non von Neumann architectures: Neuromorphic, automata….
- 3D integration of memory architectures…

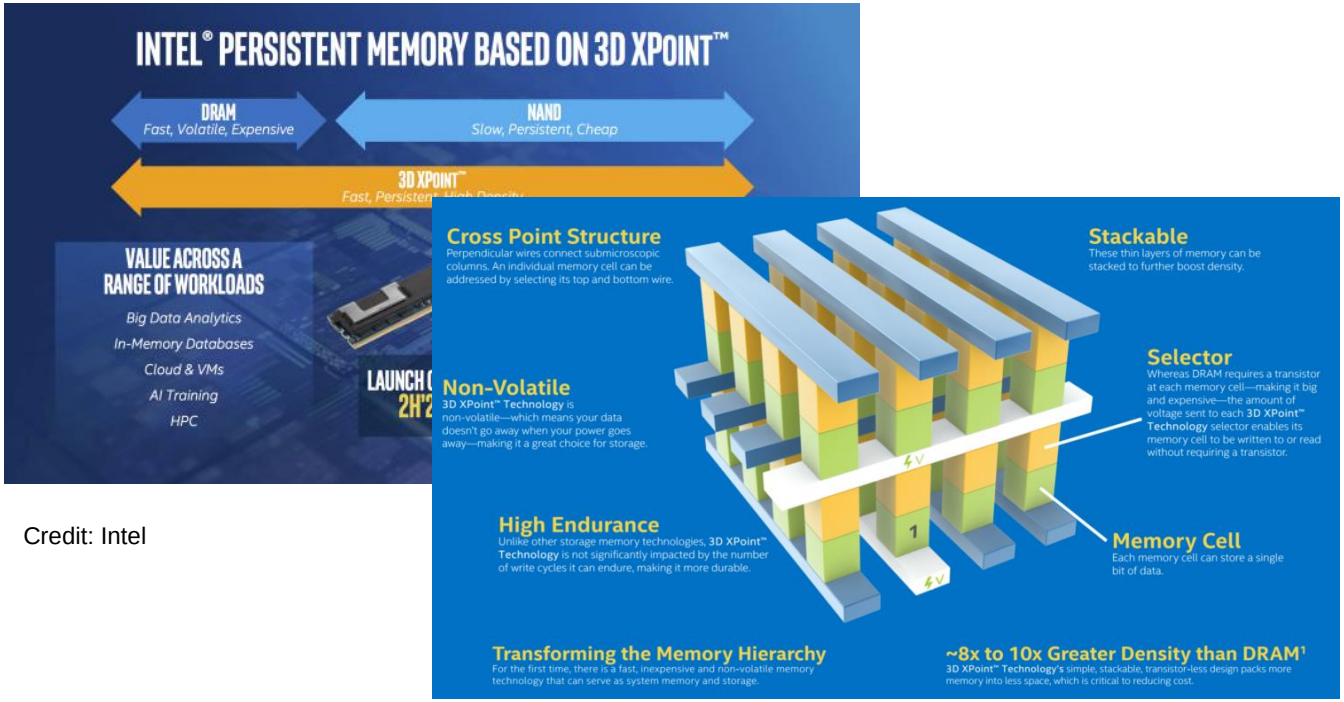






OAK RIDGE National Laboratory

3D packaging….Nonvolatile memory….Carbon Nanotubes/Graphene…



Credit: Intel

Credit: MIT

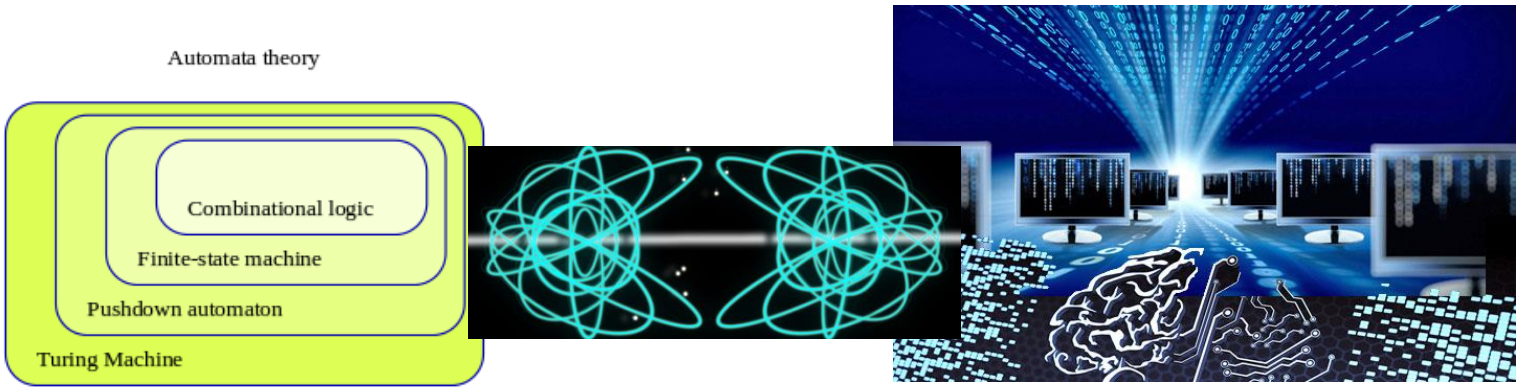
Instead of relying on silicon-based devices, researchers and Stanford University and MIT have built a new chip that uses carbon nanotubes and resistive random-access memory (RRAM) cells. The two are built vertically over one another, making a new, dense 3-D computer architecture with interleaving layers of logic and memory. This work was funded by DARPA, NSF, SRC, STARnet SONIC, and member companies of the Stanford SystemX Alliance.

Credit: Irene Qualters (NSF), Beyond CMOS Computing: The Interconnect Challenge, November, 2017

OAK RIDGE National Laboratory

Evaluating post-CMOS alternatives

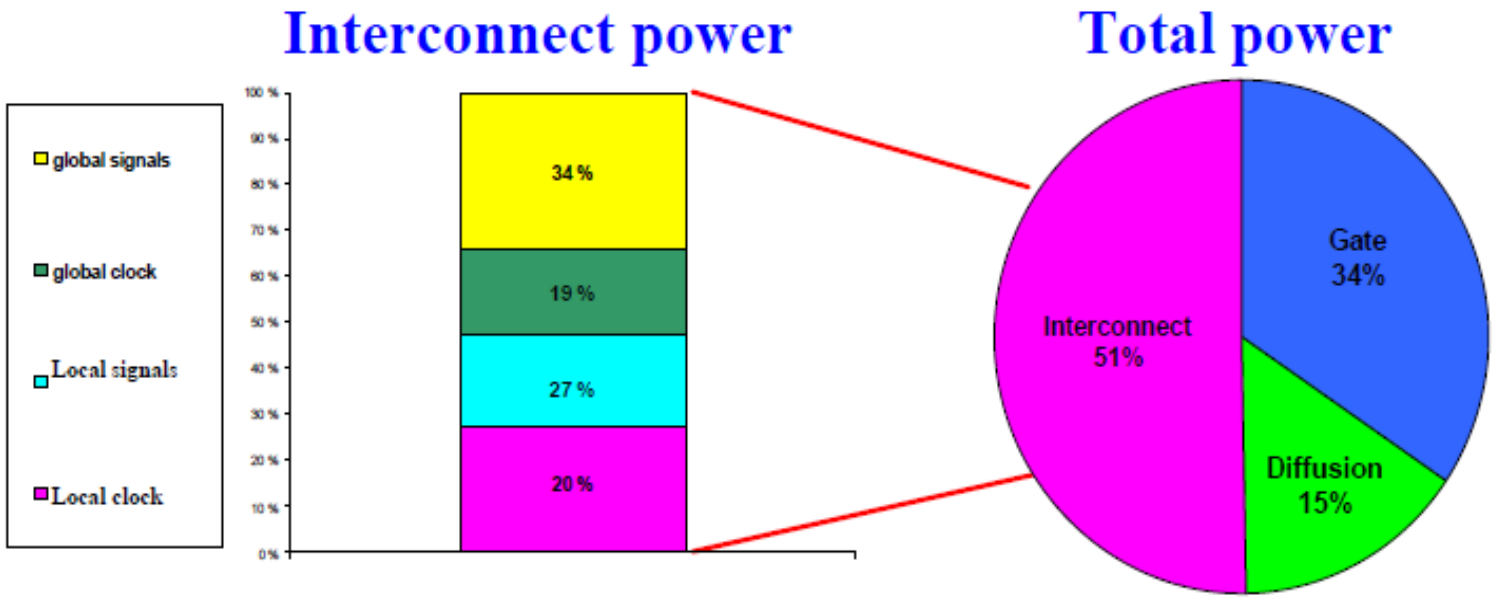
- There are no evident replacement technologies yet
- Post-CMOS alternatives need to meet
  - Scalability criteria: must allow density increases and corresponding energy reduction
  - Signal to noise immunity (e.g. quantum computing)
  - Scalable manufacturability: implementation at industrial scale (e.g. carbon based materials, carbon nanotubes)
  - Cost efficiency: some solutions based on III-V semiconductors, optical computing, are too costly
  - Large scale demonstration: need to demonstrate end to end solution (e.g. lack of large scale cryogenic memory for superconducting computing)



Beyond CMOS Computing: The Interconnect Challenge Workshop, 2017

- Keynote and two invited talks
- Four sessions
  - Benchmarks
  - Interconnects
  - Balance of the system design
  - CMOS/Beyond CMOS Integration

Interconnect technology identified as largest single factor limiting both performance and power as well as reliability



Credit: N. Magen (Intel), et al; SLIP Workshop 2004 (from A. Naeemi, Georgia Tech)

Credit: Irene Qualters (NSF), Beyond CMOS Computing: The Interconnect Challenge, November, 2017

Interconnect themes….

- Nanowire Interconnects
- Nanophotonics for Chip-scale Information Processing and Transport
- Carbon-Based Interconnects….reducing line resistance
- Silicon photonics
- SuperCables program

Beyond CMOS Computing 2017:
Energy efficient, high bandwidth digital data links between 4 and 300K

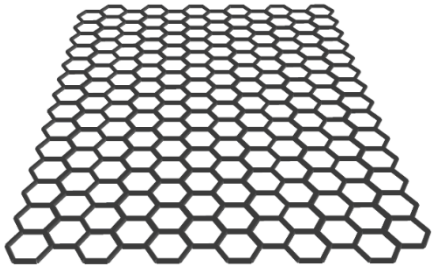
Dr Deborah Van Vechten
Office of Naval Research code 312
Superconducting Electronics Program Officer
703 696 4219
Deborah.vanvechten@navy.mil

SuperCables program

- High bit rate, energy efficient data links are essential if cryogenic forms of computing are to succeed in the Beyond CMOS world
- Electrical approaches are proven and in use, but will likely fail to scale well to large systems
- Photonic approaches are unexplored below 300K, but do multiplex well and fiber has low heat load and cross talk issues
- Work on EO components is justified to see if materials and or devices properties can enable operation in highly energy efficient manner at 4K

Graphene interconnects

Graphene



Atomically thin sheet of carbon
(thickness = 0.34 nm)

High breakdown current
(>100x Cu)

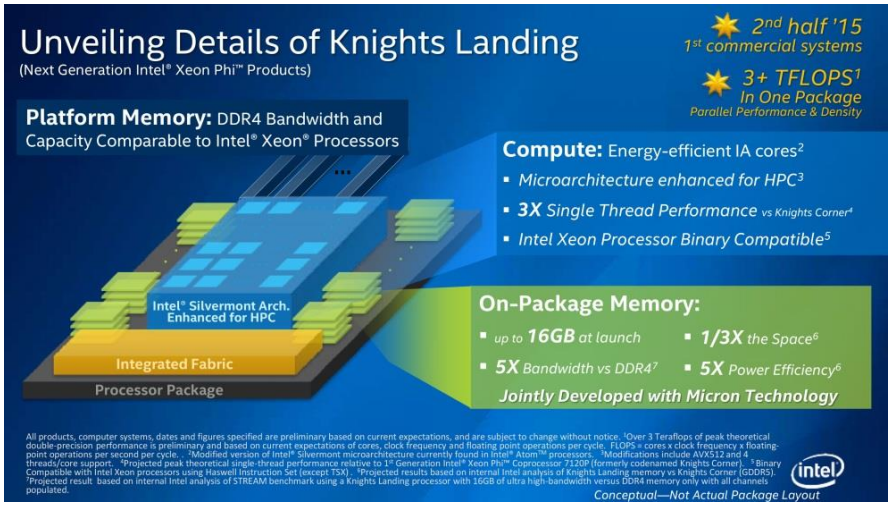
Suppressed "size effect"
($\rho_{Gr} < \rho_{Cu}$)

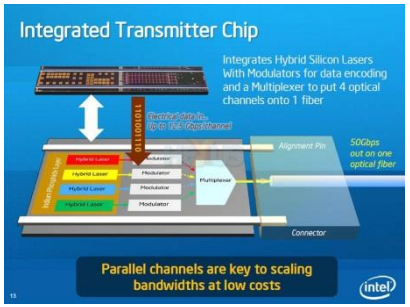
- Graphene may offer three-fold benefit:
  - Reduced size-effect at narrow linewidths ($\rho\downarrow$)
  - Improved aspect ratio ($C\downarrow$)
  - Liner reduction/elimination
- Graphene interconnects at system level:
  - Up to ~30% EDP improvement
  - 9% higher speed or 16% lower energy consumption
- Routing optimization can improve local R_{wire} impact
  - Requires place-and-route to see the effect
  - Consider utilizing area instead of performance improvement

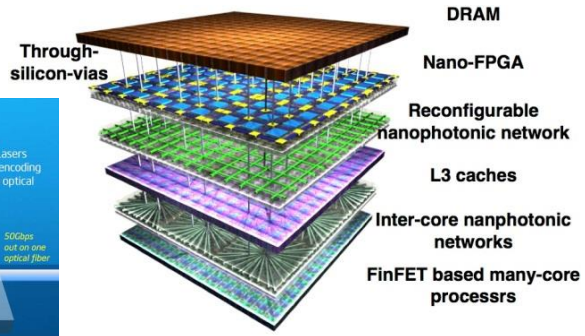
[1] L. Li, H.-S.P. Wong, *et al.*, *2016 IEDM*

Near term solutions may be Silicon based

- **Three dimensional integration and packaging**
  - Intel's Knights Landing, integrated memory, integrated fabric, parallel performance
  - Vertically integrated ecosystem: System Scalable framework
- **Silicon Photonics**
  - for fast data movement: use silicon as an optical medium, optical and electronic components are connected on a single chip







"Today, optics is a niche technology. Tomorrow, it's the mainstream of every chip that we build", --Pat Gelsinger, former Intel senior vice president , 2006.

OAK RIDGE National Laboratory

Beyond CMOS device benchmarking

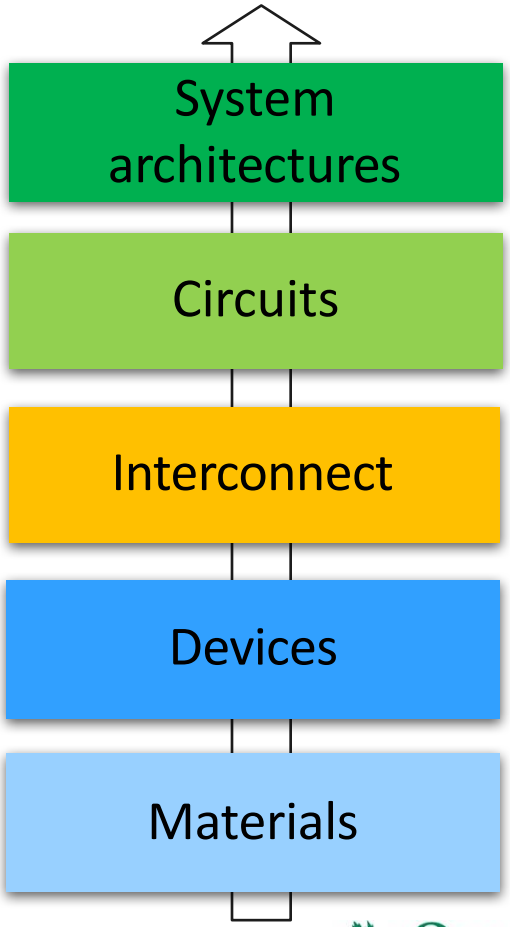
The development of SRC NRI benchmarking

Phase 1: led by Kerry Bernstein (IBM)
- Build circuit gates (inverter, NAND, adder) with NRI devices to benchmark performance (delay, power, area) against CMOS
- Consider circuit parameters, e.g., span of control, noise immunity, logical effort

Phase 2: led by Dmitri Nikonov and Ian Young (Intel)
- Develop uniform and transparent criteria and methodologies for benchmarking
- Improve quality of assumptions and device data for benchmarking

Phase 3: led by Azad Naeemi (Georgia Tech)
- Expand benchmarking to STARnet devices
- Expand benchmarking to non-Boolean logic and memory applications

Benchmarking considerations:
- Theoretical projection vs. experimental demonstration
- Accuracy and rigorousness of assumptions and parameters



Credit: An Chen (IBM/SRC NRI), Beyond CMOS Computing: The Interconnect Challenge, November, 2017

Benchmarking Beyond CMOS devices

Lessons Learned from the SRC NRI Program

- No beyond-CMOS device has been proven to be capable of replacing CMOS for Boolean logic and von Neumann architectures.
  - If there is a CMOS solution, the difficulty and requirement of adopting a beyond-CMOS technology will be significantly higher.
  - There are still interesting low-power device mechanisms worth exploring
  - Beyond-CMOS devices may find promising opportunities for novel computing paradigms beyond Boolean logic and von Neumann architectures.
- Novel material and processing are essential for beyond-CMOS research.
  - There is significant material and processing barrier for most beyond-CMOS devices.
  - Need to focus and address the hard problem.
- A holistic approach is needed to address new beyond-CMOS research directions.
  - Benchmarking is essential and needs to start from the beginning.
- Public-private collaboration is critical for funding the basic research needed to advance the semiconductor and computing technologies.

Credit: An Chen (IBM/SRC NRI), Beyond CMOS Computing: The Interconnect Challenge, November, 2017